2009 NIST Language Recognition Evaluation Evaluation Overview

Craig Greenberg
Alvin Martin

Based on the NIST presentation at:

LRE09 Workshop
Baltimore, Maryland, USA
June 24-25, 2009

Outline

- Evaluation Overview
- Participants
- Overall Evaluation Results
- Performance History
- Performance by Language
- Performance by Data Type
- Summary

What's New For LRE09?

- Primary (new) data is broadcast telephone bandwidth Voice of America (VOA) data
  - Early analysis of VOA data done at Brno
  - Collected and audited by the LDC
  - Large VOA corpora and designated segments made available for development in languages for which previous LRE conversational telephone speech (CTS) data not available
- 23 target languages, 16 out-of-set languages
- Larger numbers of test segments available for most languages
- Segments of approximately 3, 10, or 30 seconds of speech all grouped together (but performance examined separately)
  - Careful listening to 10 and 3 second CTS segments
  - Found overlapping 10 and 3 second CTS speech segments that minimized time elapsed
  - Selected 10 and 3 second VOA by iterating over each sample and:
    - Let Eavg_i be the average energy in candidate segment seg_i
    - Let Emax be the maximum of Eavg_i over all seg_i
    - Let score_i be the score for segment seg_i, with score_i = max(Ew1, Ew2, .05*Emax)/Eavg_i
    - Chose the seg_i that minimizes score_i.
  - Feather-cut voa segments using 10ms linear ramp

LRE09 Languages

(counts are for 30-second segments)

Lang.	VOA Train	VOA Test	CTS Test
Amharic	171	398	-----
Bosnian	194	355	-----
Cantonese	-----	62	316
Creole-Haitian	186	323	-----
Croatian	181	376	-----
Dari	194	389	-----
English-Am.	-----	374	522
English-Ind.	-----	-----	574
Farsi	-----	338	52
French	196	395	-----
Georgian	142	399	-----
Hausa	200	389	-----
Hindi	-----	397	270
Korean	-----	318	145
Mandarin	-----	390	625
Pashto	197	395	-----
Portuguese	166	397	-----
Russian	-----	254	257
Spanish	-----	385	-----

Lang.	VOA Train	VOA Test	CTS Test
Turkish	194	394	-----
Ukrainian	194	388	-----
Urdu	-----	347	32
Vietnamese	-----	27	288
Arabic	*Out-of-set*	*187*	-----
Azerbaijani	*Out-of-set*	*366*	-----
Belorussian	*Out-of-set*	*363*	-----
Bengali	*Out-of-set*	-----	*43*
Bulgarian	*Out-of-set*	*375*	-----
Italian	*Out-of-set*	-----	*30*
Japanese	*Out-of-set*	-----	*180*
Punjabi	*Out-of-set*	-----	*9*
Romanian	*Out-of-set*	*400*	-----
Shanghai-Wu	*Out-of-set*	-----	*69*
Southern-min	*Out-of-set*	-----	*48*
Swahili	*Out-of-set*	*396*	-----
Tagalog	*Out-of-set*	-----	*84*
Thai	*Out-of-set*	-----	*188*
Tibetan	*Out-of-set*	*368*	-----
Uzbek	*Out-of-set*	*382*	-----

Test Conditions

- **Closed-set:** segment languages are limited to in-set languages, all (in-set) target languages
- **Open-set:** segment languages also include (undisclosed) out-of-set languages
- **Language pairs:** Segment and target languages limited to two, for each possible in-set pair
  - Thus always a single alternative hypothesis for each trial
  - Certain pairs designated as of particular interest

Cantonese -- Mandarin	Hindi -- Urdu
Portuguese -- Spanish	Farsi -- Dari
Creole -- French	Bosnian -- Croatian
Russian -- Ukrainian	Engl. (American) – Eng. (Indian)

System Input/Output

- Input: all trials for a test condition, consisting of all pairings of a test segment and a target language/dialect
- Output: for each trial
  - a decision (true/false)
  - a score on which the decision is based, where higher scores imply greater belief that "true" is the correct decision
    - Systems were asked to specify if their scores could be interpreted as log-likelihood ratios (llr's):

$$= \ln P(\text{data} \mid \text{target language i}) - \ln P(\text{data} \mid \text{not target language i})$$

where ln is the natural logarithm function

Evaluation Rules

- All 41793 test segments of all durations must be processed for each target language
- Each test segment must be processed separately and without any knowledge of other test segments.
  - Normalization over multiple test segments is *NOT* allowed.
- Side knowledge of the sex or other characteristics of the test speaker is *NOT* allowed.
  - Unless obtained by automatic means.
- Listening to the evaluation data or any other experimental interaction with the data is *NOT* allowed before all test results have been submitted.
- Use of knowledge of the full set of target languages/dialects for each test *IS* allowed.

Basic Performance Measure

$$C(L_T, L_N) = C_{\text{Miss}} \cdot P_{\text{Target}} \cdot P_{Miss}(L_T) + C_{\text{FA}} \cdot (1 - P_{\text{Target}}) \cdot P_{FA}(L_T, L_N)$$

where

L_T and L_N are a target/non-target language pair

C_{Miss}, C_{FA} and P_{Target} are application model parameters

For LRE09, the application parameters will be:

$C_{\text{Miss}} = C_{\text{FA}} = 1$, and

$P_{\text{Target}} = 0.5$

Average Performance

$$C_{avg} = \frac{1}{\mathrm{N_L}} \cdot \sum_{L_T} \left\{ \begin{array}{l} \mathrm{C_{Miss}} \cdot \mathrm{P_{Target}} \cdot P_{Miss}(L_T) \\ + \sum_{L_N} \mathrm{C_{FA}} \cdot \mathrm{P_{Non-Target}} \cdot P_{FA}(L_T, L_N) \\ + \mathrm{C_{FA}} \cdot \mathrm{P_{Out\text{-}of\text{-}Set}} \cdot P_{FA}(L_T, L_O) \end{array} \right\}$$

where

N_L is the number of languages in the (closed-set) test

L_O is the Out-of-Set "language"

$$\mathrm{P_{Out\text{-}of\text{-}Set}} = \left\{ \begin{array}{l} 0.0 \text{ for the closed - set condition} \\ 0.2 \text{ for the open - set condition} \end{array} \right.$$

and

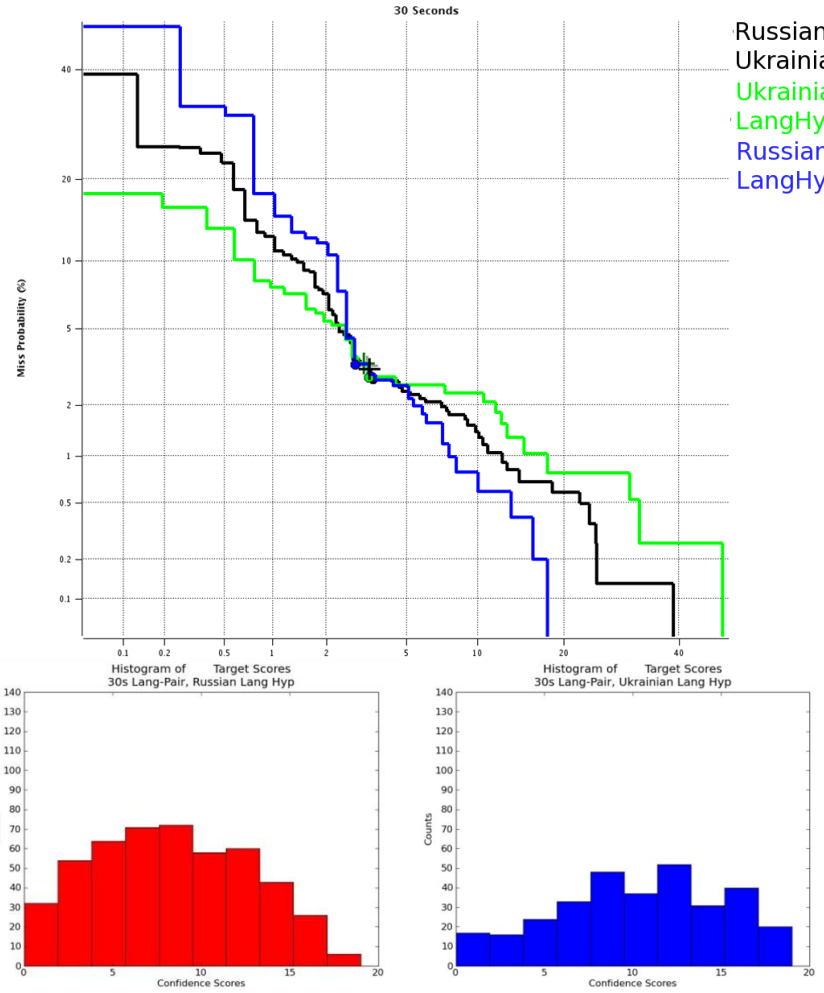
$$\mathrm{P_{Non-Target}} = \left(1 - \mathrm{P_{Target}} - \mathrm{P_{Out\text{-}of\text{-}Set}}\right) / \left(\mathrm{N_L} - 1\right)$$

DET Curves

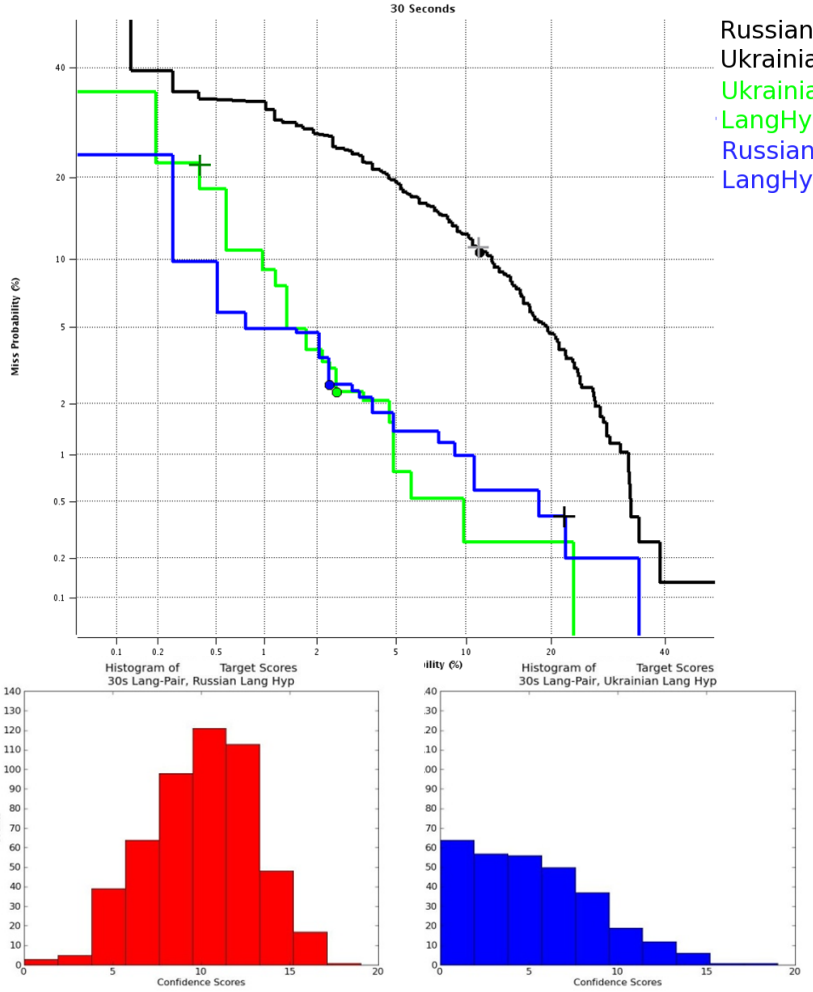
- In speaker recognition all trials are pooled to create the DET curve
- In language recognition DET's are computed separately for each language pair and then:
  - DET's are averaged across all non-target languages to produce a DET for each target language
  - DET's for all target languages are averaged to produce an overall DET
- The quality of calibration across languages affects the overall multi-target language DET curves
  - This is illustrated dramatically for the language-pair case
    - the DET's for the two single targets should be symmetric
    - these two DET's should have the same EER.
    - but if the scores are not properly calibrated the combined DET will be degraded
    - the next slide shows an example

Russian-Ukrainian Pair Example

System-1



System-2



Outline

- Evaluation Overview
- Participants
- Overall Evaluation Results
- Performance History
- Performance by Language
- Performance by Data Type
- Summary

Participating Sites/Teams (1)

System Name	Site	Location
ATVS	Universidad Autonoma de Madrid	Madrid, Spain
BUT-AGN	Brno University of Technology Agnitio	Brno, Czech Republic Somerset West, South Africa
CASIA	Institute of Automation, Chinese Academy of Sciences	Beijing, China
CUHK	Chinese University of Hong Kong	N.T., Hong Kong
EHU	University of the Basque Country	Bizkaia, Spain
IFLY	iFlyTek Speech Lab, EEIS University of Science and Technology of China	HeFei, AnHui, China
IIR	Institute for Infocomm Research	Singapore
IOA	Institute of Acoustics, Chinese Academy of Sciences	Beijing, China
L2F	L2F-Spoken Language Systems Lab INESC-ID Lisboa	Lisbon, Portugal
LIA	Laboratorie Informatique D'Avignon	Avignon, France

Participating Sites/Teams (2)

System Name	Site	Location
LIMSI	CNRS-LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur)	Orsay, France
LPT	Loquendo Politecnico di Torino	Torino, Italy Torino, Italy
MIT	MIT Lincoln Laboratory	Lexington, MA, USA
NTUT	National Taipei University of Technology, Department of Electrical Engineering & Graduate Institute of Computer and Communication Engineering	Taipei, Taiwan
THU	Tsinghua University Department of Electrical Engineering	Beijing, China
TNO	Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek	Soestenberg, The Netherlands

Outline

- Evaluation Overview
- Participants
- Overall Evaluation Results
- Performance History
- Performance by Language
- Performance by Data Type
- Summary

Overall Evaluation Results

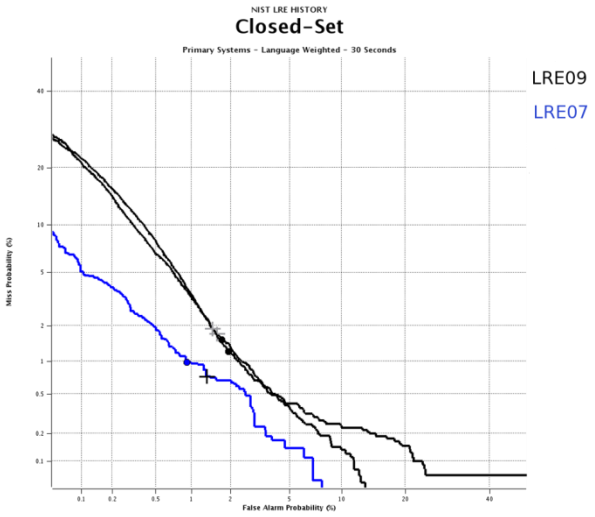
See web page summary:

http://www.itl.nist.gov/iad/mig/tests/lre/2009/lre09_eval_results/index.html

Outline

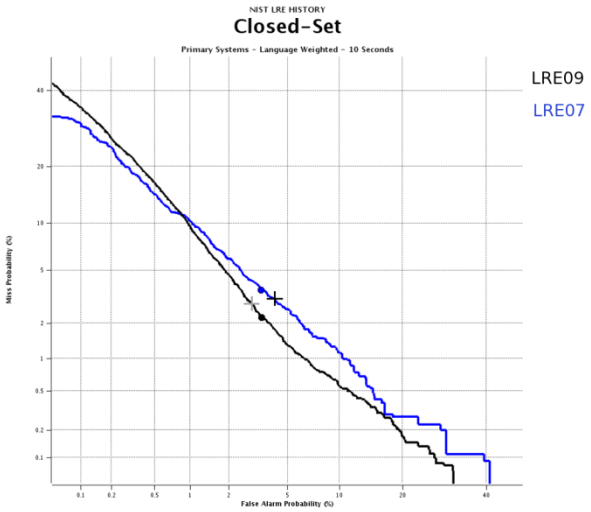
- Evaluation Overview
- Participants
- Overall Evaluation Results
- Performance History
- Performance by Language
- Performance by Data Type
- Summary

Best System – Closed Set
2007, 2009

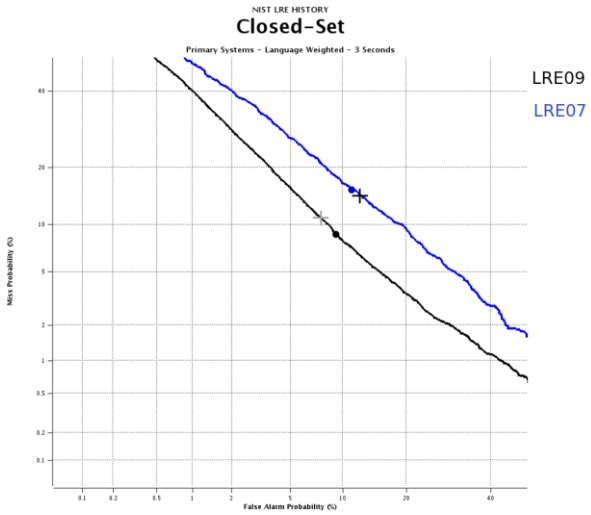


30sec

- Co-winners in 30 sec trials
- Performance loss in 30 sec trials compared with LRE07



10sec

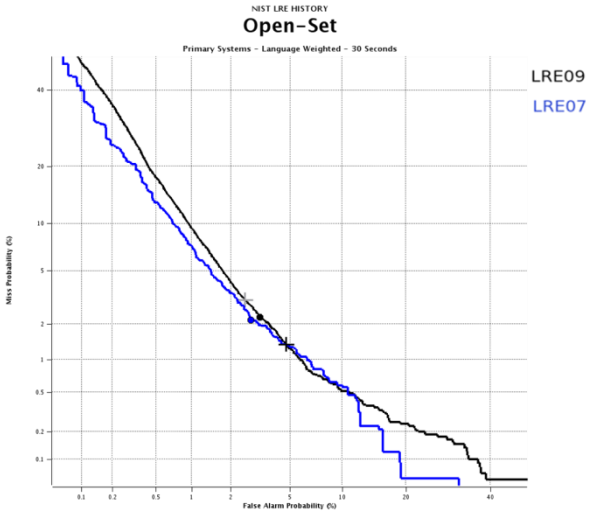


3sec

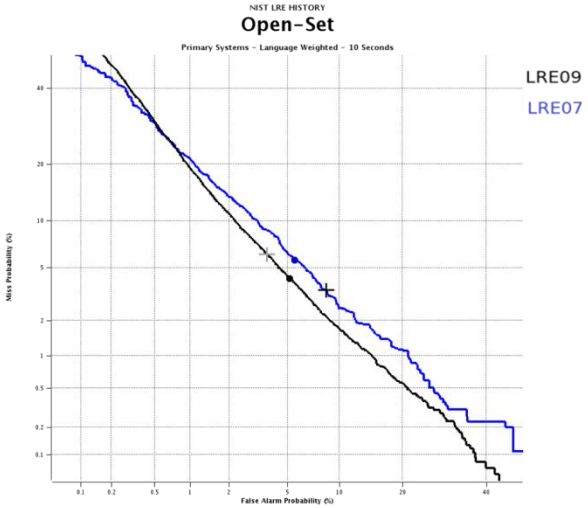
- 3 sec saw better performance compared with LRE07
- Improved selection of 3 sec segments

Best System – Open Set

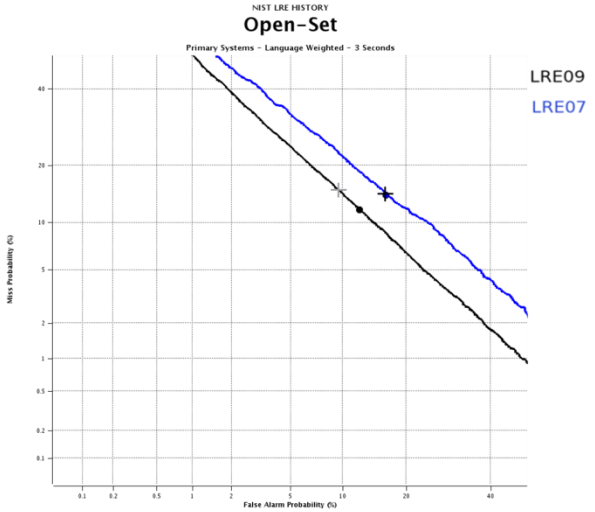
2007, 2009



30sec



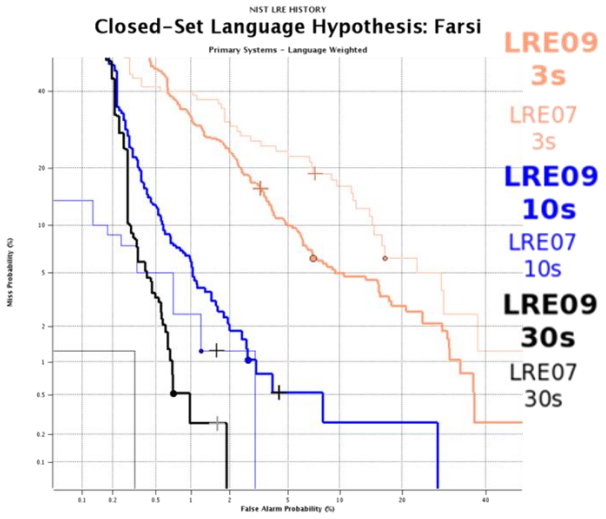
10sec

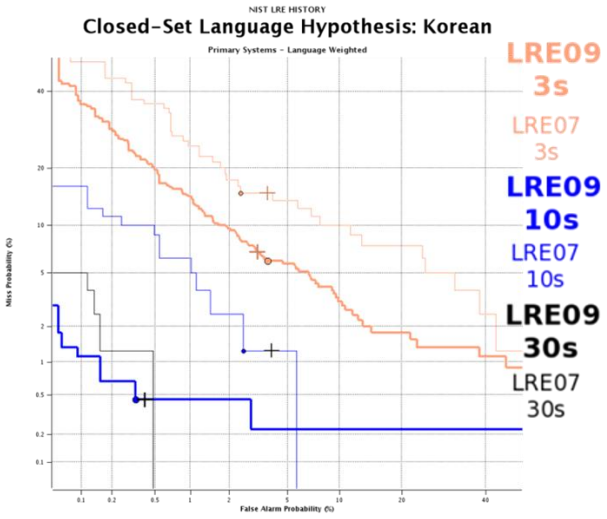


3sec

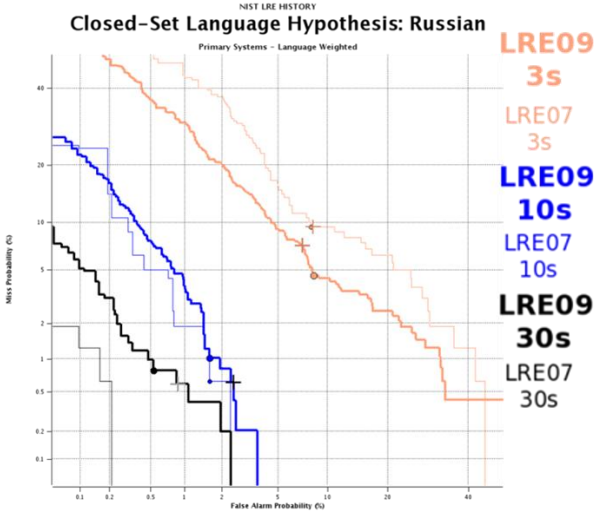
Best Systems by Target Language

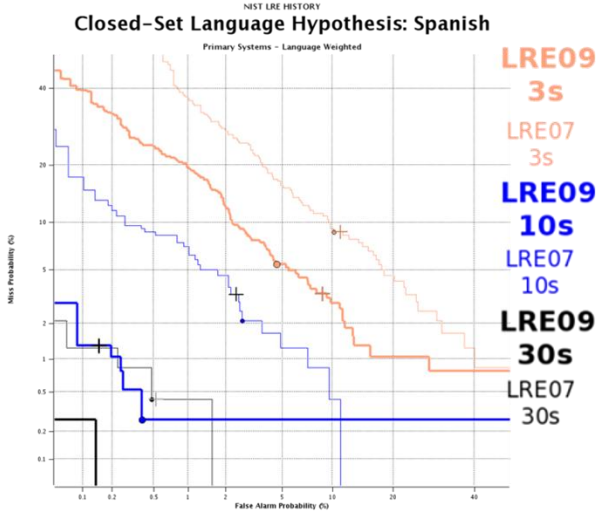
Closed-Set – 2007, 2009

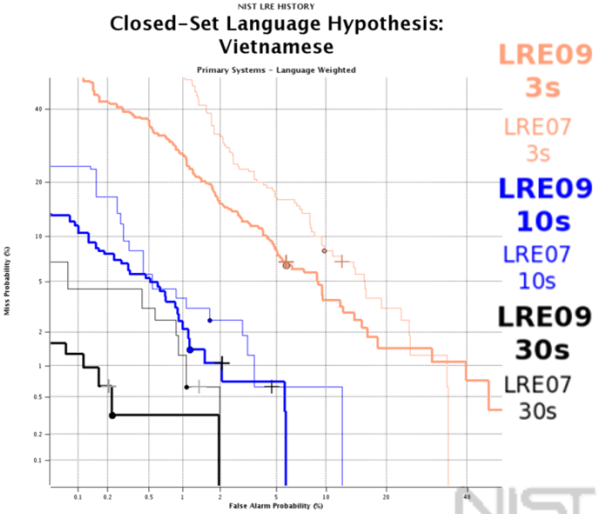




30 sec Korean off chart!

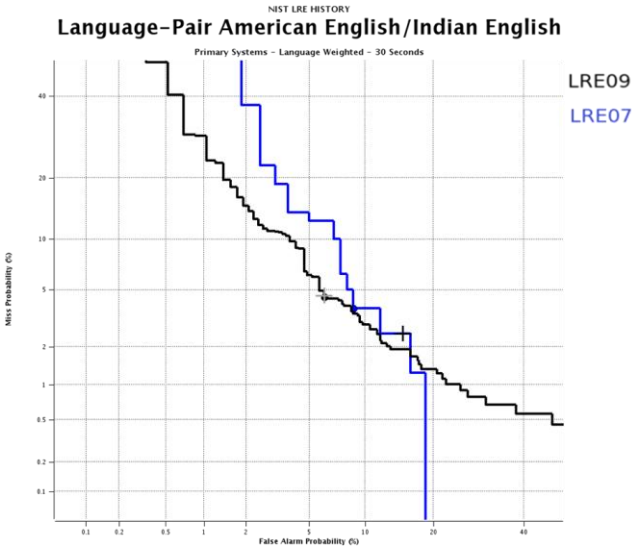




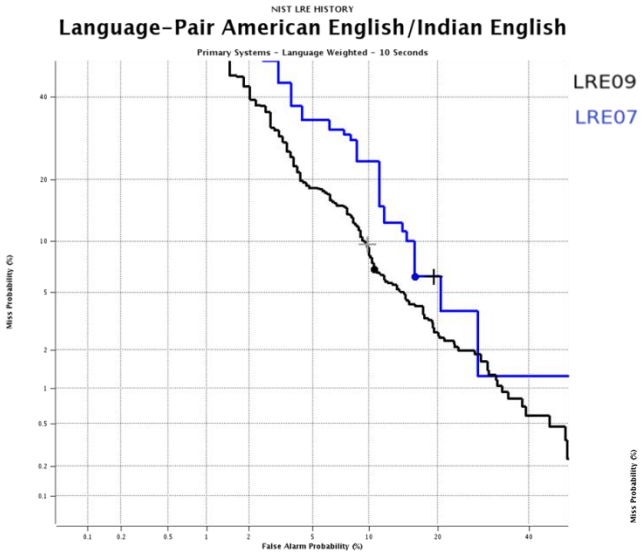


Best System - Recognizing American English for American English/Indian English Language Pair

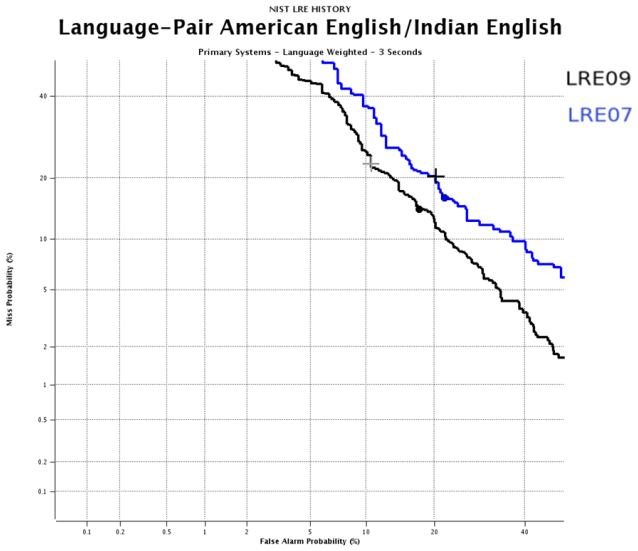
2007, 2009



30sec



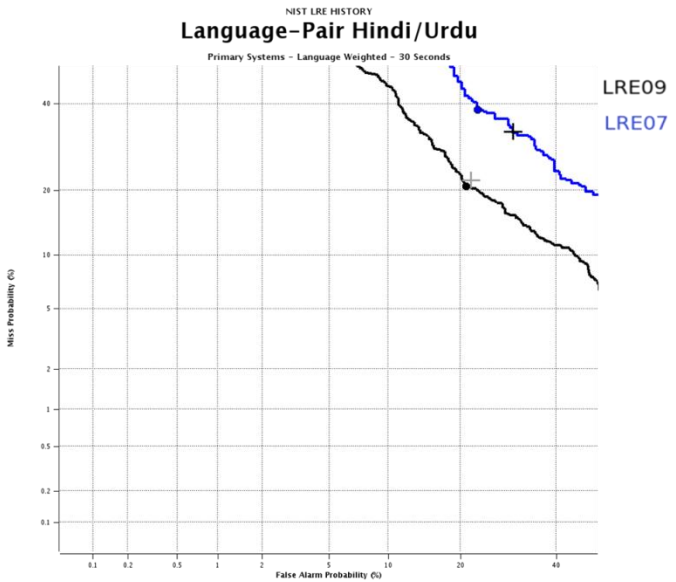
10sec

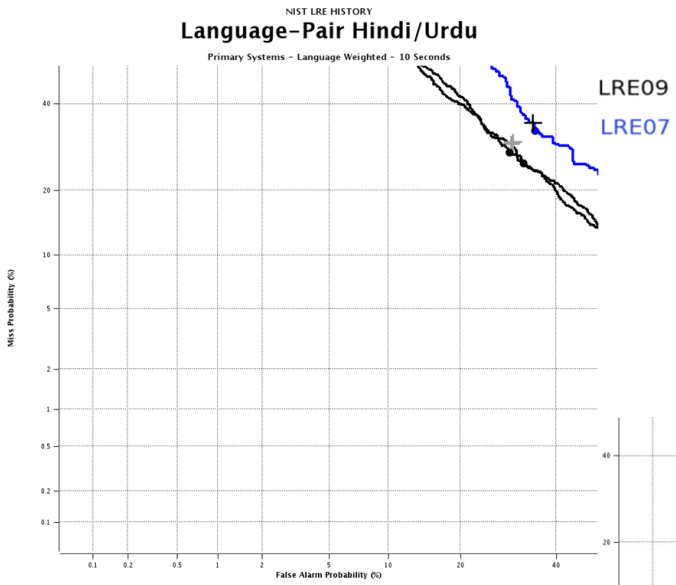


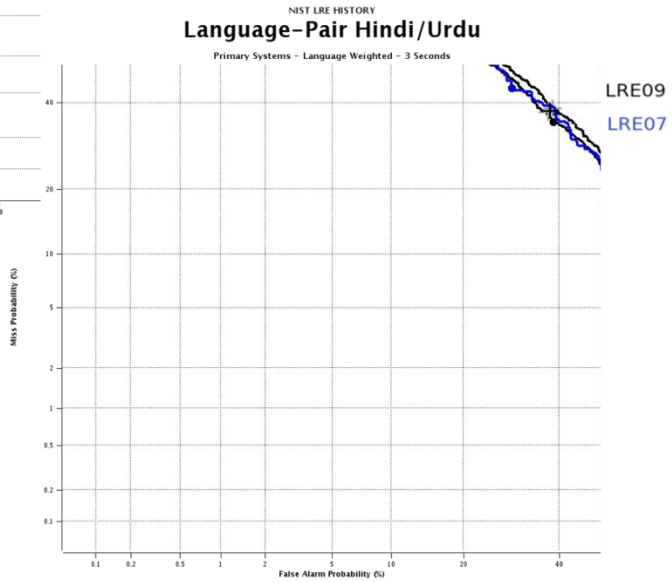
3sec

- Improvement for all three durations

Best System - Recognizing Hindi for Hindi/Urdu Pair 2007, 2009







- Real improvement in 30sec and 10sec

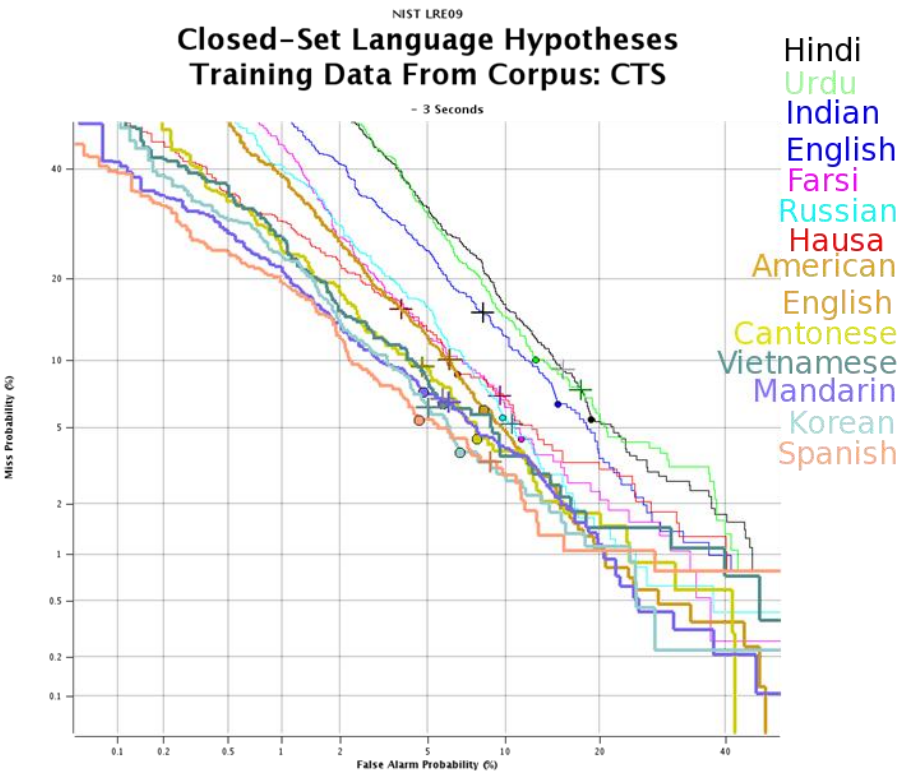
- 3 sec still challenging

Outline

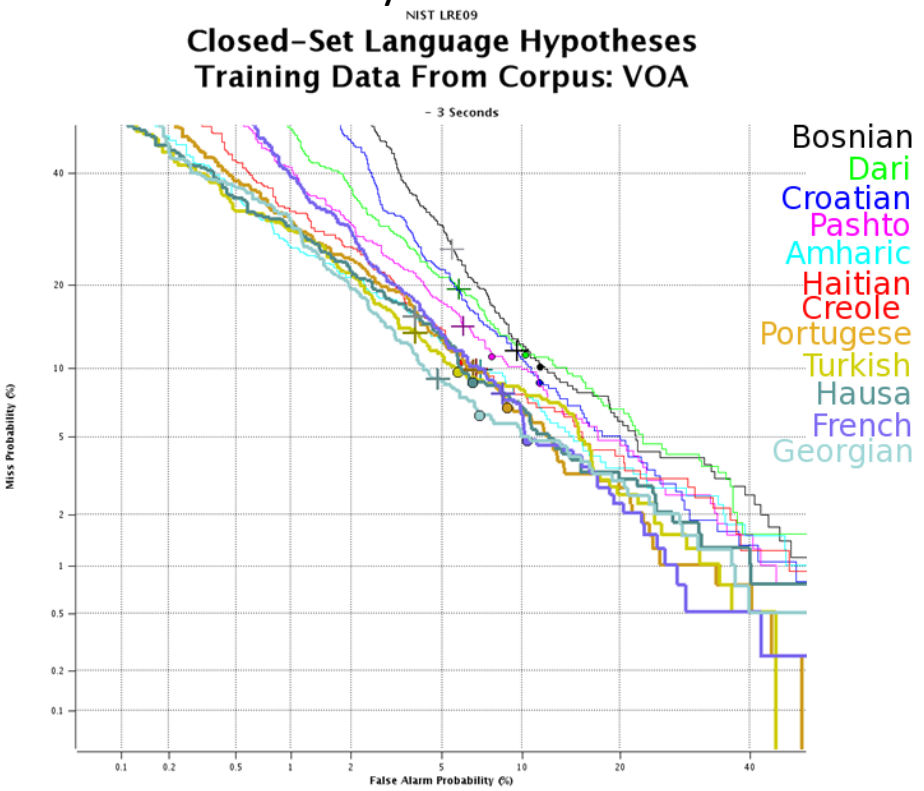
- Evaluation Overview
- Participants
- Overall Evaluation Results
- Performance History
- Performance by Language
- Performance by Data Type
- Summary

Closed Set Performance by Target Language

System 1



System 2



- Indian languages were challenging
- CTS training somewhat better performance

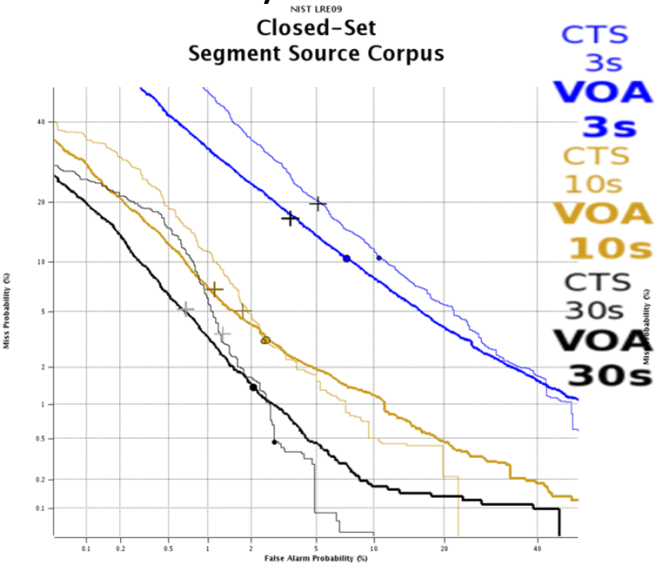
Outline

- Evaluation Overview
- Participants
- Overall Evaluation Results
- Performance History
- Performance by Language
- Performance by Data Type
- Summary

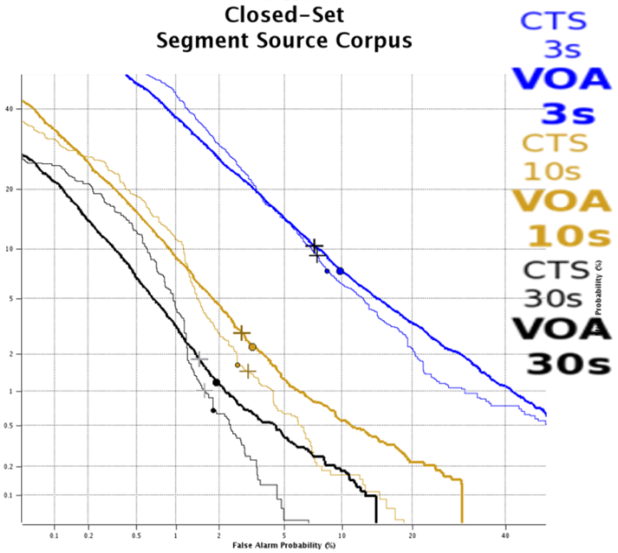
Closed Set Performance by Data Type

- VOA and CTS performance broadly comparable
- CTS curves less linear, with better performance at high FA rates

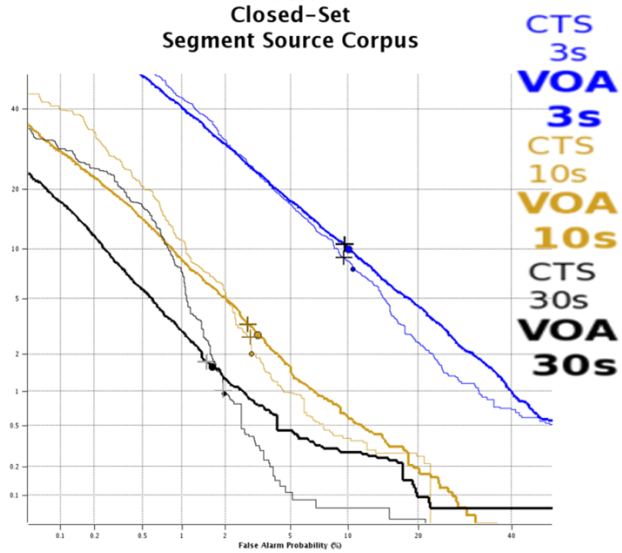
System 1



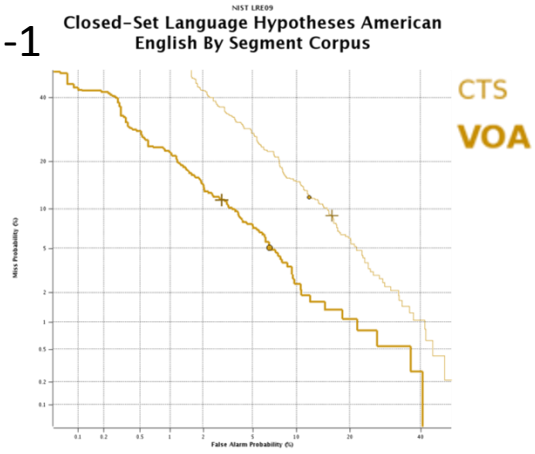
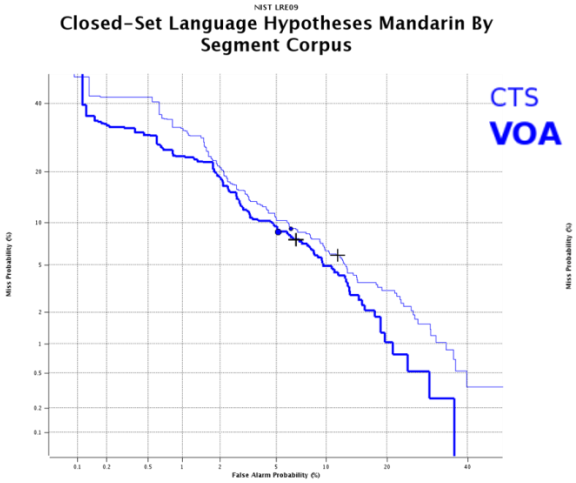
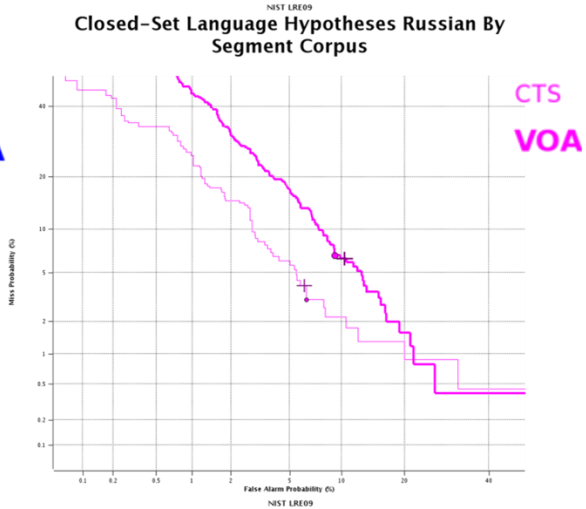
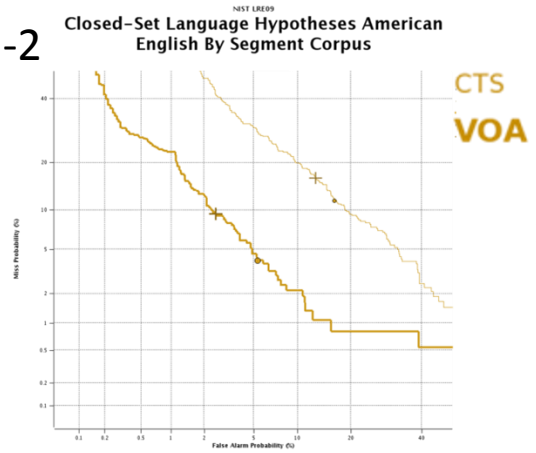
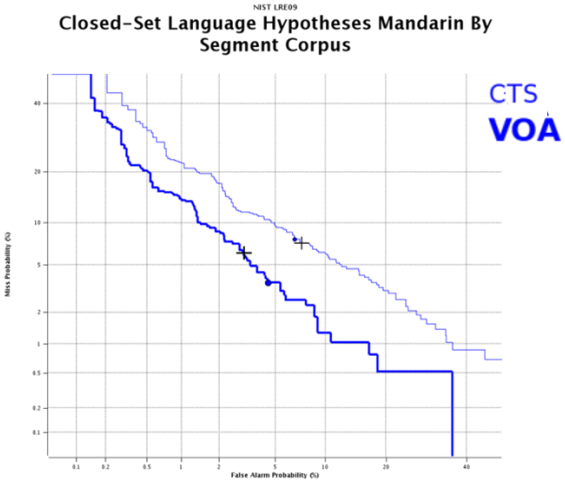
System 2



System 3



Single Target Language Performance by Data Type (3sec)

	English	Mandarin	Russian
System-1			
System-2			

Summary and Issues

- LRE09 was essentially successfully conducted largely utilizing narrowband broadcast speech
  - Performance on VOA was comparable to that with CTS
  - Larger numbers of test segments were included
  - But speakers were often repeated
- Some performance improvement seen compared with LRE07, particularly for shorter duration segments
- Similar (particularly mutually comprehensible) languages present performance (and auditing) challenges
- Some issues with scoring and DET curves
  - Should language pairs be emphasized?
  - Does LRE09 provide a model for future evaluations?